# The Physics of AI

SCOTT BARNETT, ALEKSANDAR PASQUINI, STEFANUS KURNIAWAN, SHANGEETHA SIVASOTHY, RHYS HILL, RAJESH VASA, Deakin University, Australia

Generative AI systems are governed by a set of structural regularities, what we term the Physics of AI, that define how they behave in practice. These regularities are not optional design choices but unavoidable constraints. Through three case studies, a retrieval-augmented manual assistant, a semantic document-matching service, and a test-data generation tool, we examine how these physical constraints manifest in real projects and identify ten recurring challenges. By interpreting these findings through the lens of the Physics of AI, we propose five actionable practices that make AI engineering more reproducible, interpretable, and accountable. While unavoidable, we believe that accounting for the Physics of AI changes engineering methodology and practices, not necessarily the outcome.

CCS Concepts: • **Software and its engineering → Software creation and management**.

Additional Key Words and Phrases: Generative AI systems, SE4AI, AI engineering, Case Study



## 1 Introduction

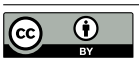
Artificial intelligence systems increasingly underpin critical infrastructure, decision-making, and creative processes. Yet, despite rapid advances in capability, the engineering of AI remains dominated by uncertainty, brittleness, and emergent behaviour. Models trained under one distribution fail under another, evaluation benchmarks lose relevance as tasks evolve and explanations that seem coherent conceal hidden dependencies in data and context [17, 21, 22, 27]. These persistent challenges reveal that these problems are not simply design flaws but structural features of AI systems themselves.

We call these structural features the Physics of AI. Much like how physical laws govern motion and energy, these principles govern uncertainty, drift, entanglement, and resource limits within AI systems. They define what can and cannot be engineered away, shaping the boundaries of reliability, interpretability, and control. Recognising these forces allows practitioners to design within the inherent limits of AI.

Current literature attempts to address robustness [13], interpretability [16], ethics [20], or data drift [18] in isolation. This leads to brittle, failure-prone systems [14, 19]. Only by understanding how these limits interact holistically can we design resilient, trustworthy AI. Additionally, as AI-infused systems become more complex and critical, stakeholders need a shared vocabulary and set of principles from a unified conceptual framework. The Physics of AI provides such a foundation by articulating what can be achieved, what must be measured, and where tradeoffs are immutable.

We derived these Physics of AI principles through a synthesis of existing literature and insights gained from our empirical work. From this analysis, we identify nine regularities that recur across model development, deployment, and evaluation. Each regularity represents a structural constraint that persists regardless of model type, data source, or domain. We then apply this framework to three diverse case studies spanning retrieval-augmented generation, semantic document matching, and automated data synthesis. Through these, we demonstrate how the Physics of AI manifests in practice and distil five actionable design principles that help engineers build systems that are more observable, adaptive, and accountable under real-world constraints.

Author's Contact Information: Scott Barnett, Aleksandar Pasquini, Stefanus Kurniawan, Shangeetha Sivasothy, Rhys Hill, Rajesh Vasa, {scott.barnett,aleksandar.pasquini,stefanus.kurniawan,s.sivasothy,rhys.hill,rajesh.vasa}@deakin.edu.au, Deakin University, Applied Artificial Intelligence Initiative, Geelong, Australia.



Contributions arising from this work include:

- A conceptual framework called the Physics of AI, that defines a set of structural regularities governing how AI systems behave in practice.
- An application of the Physics of AI lens to a catalogue of empirical findings from three generative AI case studies.
- A set of actions that incorporate the principles of the Physics of AI and our empirical findings, providing guidance for the engineering of generative AI systems.

## 2 Related Work

The recognition that computation is governed by physical limits, where error, uncertainty, and resource expenditure are unavoidable, originates in the foundational work of Bennett and Landauer (1985) [4]. Continuing on from these insights, modern research in machine learning and trustworthy AI has sought to articulate system-level desiderata for robustness, traceability, and operational resilience [1, 2]. These include rigorous versioning, provenance tracking, and continuous monitoring, practices now viewed as essential for maintaining reliability in adaptive systems. Advances in representation learning further highlight the importance of causal structure and interpretability grounded in empirical data and physical plausibility [26].

Extending this line of thought, studies on distribution shift and model drift reveal an inherent limit to AI performance when systems are not continuously retrained and recalibrated. Koch et al. (2024) show that medical AI systems operate under persistent environmental and data distribution changes, rendering static validation insufficient [15]. Subsequent analyses similarly argue that robustness and resilience depend on engineered mechanisms for adaptation and lifecycle provenance tracking [5].

A further structural limit lies in alignment. Overreliance on narrow performance metrics creates persistent gaps between model behaviour and human or societal intent, motivating the move toward multi-dimensional, context-aware evaluation frameworks [24]. Parallel developments in AI ethics and regulation echo this need. Emerging standards and legislative proposals increasingly demand empirically verifiable transparency, explainability, and observability to support accountability for error, bias, and drift [12].

While the desiderata and regulatory literature provide valuable modular checklists and procedural safeguards, they treat error, drift, feedback, and alignment as separate concerns. Few integrate them into a unified conceptual framework that connects technical, organisational, and societal domains. Our Physics of AI framework advances this goal in two ways:

(1) We argue that these constraints should be unified as principles across technical, operational, and societal boundaries rather than listing them separately as desiderata or via regulatory procedures. Such integration enables their use early in the design process, where they can most effectively guide system architecture and evaluation.
(2) We frame the constraints as "hard limits" or first principles, fundamental like physical laws. Hence, Physics of AI provides a more pro-active framework for understanding and designing AI.

## 3 The Physics of AI

We use the term Physics of AI to denote a set of persistent regularities that shape how AI systems behave in practice. These regularities are not optional design choices but structural constraints that any engineering approach must design around. We group these constraints into three categories: Adaptation, Regulation and Resilience.

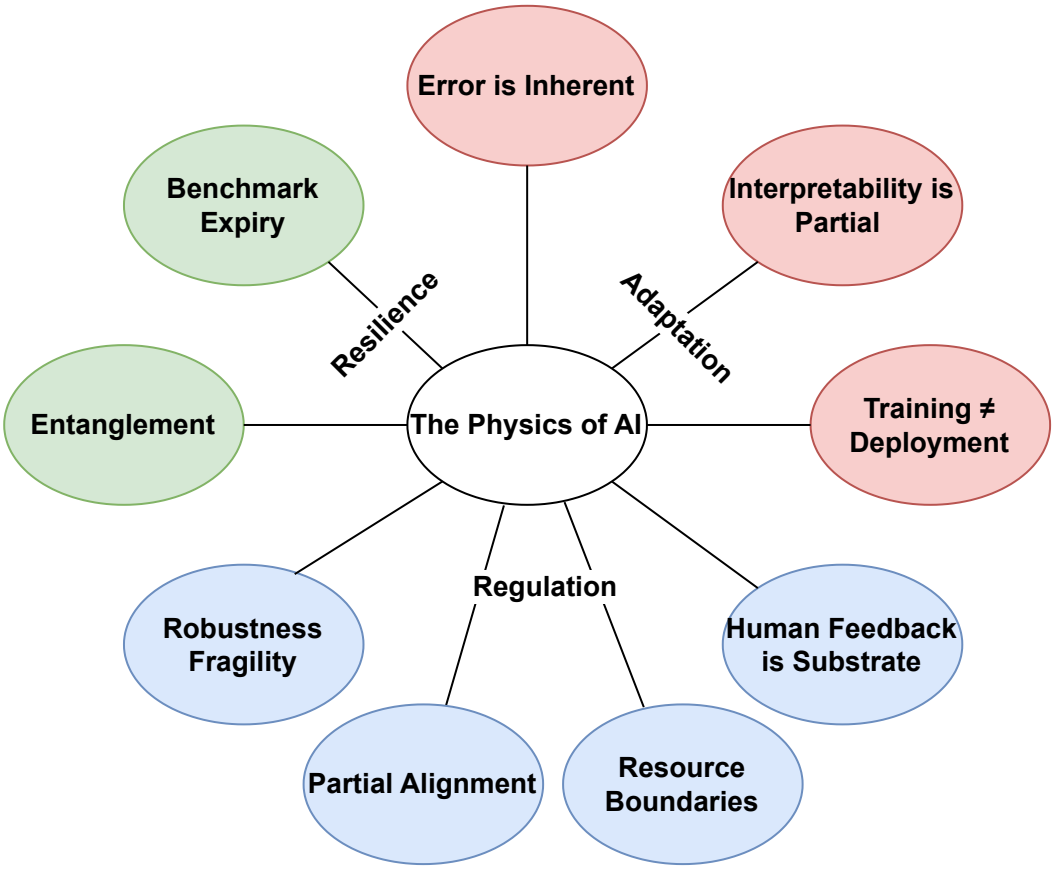



Fig. 1. There are 9 constraints that limit the ability of AI systems.

### 3.1 Adaptation Constraints

Adaptation constraints are focused on how AI systems learn and change in response to evolving environments. AI systems risk either stagnation or runaway self-amplification, where uncorrected feedback loops lead to performative drift and degraded reliability. The constraints associated with adaptation are:

**Error is inherent.** All models approximate reality; even with infinite data and ideal training, there remains uncertainty that cannot be eliminated due to randomness or unmodelled factors [3, 11]. Managing error means designing mechanisms to quantify, communicate, and tolerate imperfection rather than eliminating it.

**Interpretability is partial.** AI models operate in high-dimensional abstract feature spaces that defy simple human explanation. As a result, any explanation of a model's decision is usually a simplification or abstraction of what's really going on inside the model [17]. Thus, transparency requires logging and observability rather than totalising narratives.

**Training ≠ Deployment.** The environments in which models operate differ from those they were trained in. A model's historical training distribution can become no longer reflective of current

reality [22]. Real-world variation, through data drift, API limits, or infrastructure churn, makes design-time evaluation essential.

### 3.2 Regulation Constraints

Regulation constraints defines how AI behaviour is bounded, directed, and aligned with intended objectives. Neglecting regulation leads to misaligned optimisation as well as ethical breaches, governance lapses, and uncontrolled cost or time overruns. The key constraints underlying regulation are:

**Human feedback is substrate.** AI systems do not define their own objectives or correctness; humans do. Meaning originates from people. An AI model, at its core, is pattern-matching to proxies of human-defined concepts. AI performance depends on ongoing human calibration, which is achieved through labelling, review, and hybrid evaluation loops.

**Resource boundaries.** Quotas, latency, and cost shape feasible architectures. Cottier et al. (2024) [6] estimate that the cost of training the largest LLMs has grown by about two and a half times each year since 2016. These costs will be passed onto the users. Requirements engineering must expose the resource limits early to ensure sustainable system design.

**Partial alignment.** Quantitative metrics capture only fragments of intent. Once an AI system is explicitly trained to maximize a particular metric, that metric loses its correlation with the underlying intent [7, 8]. Effective evaluation requires complementing them with qualitative, stakeholder-grounded criteria.

**Robustness fragility.** Seemingly minor changes in configuration, dependency, or dataset can cause large behavioural shifts. For instance, subtle changes to an input that are imperceptible to a human observer, can cause a neural network to produce a highly confident misclassification [10]. Full versioning and dependency traceability are required to ensure stability.

### 3.3 Resilience Constraints

Resilience constraints addresses the system's capacity to withstand shocks, recover from failures, and sustain function under stress. Without sufficient resilience measures, systems become fragile, vulnerable to cascading failures, and experience prolonged downtime when disrupted, undermining both trust and operational continuity. Resilience in AI systems is shaped by the following constraints:

**Entanglement.** Data, code, and humans form interdependent loops. As Scully et al. (2015) stated, "Machine learning systems mix signals together, entangling them and making isolation of improvements impossible ... We refer to this as the CACE principle: Changing Anything Changes Everything" [22]. Therefore, capturing data pipelines, preprocessing steps, hyperparameters, and even the infrastructure (hardware, libraries) is crucial for understanding causality and accountability.

**Benchmark expiry.** Static benchmarks decay as tasks, data, and expectations evolve [21]. For instance, dialogue evaluation has shifted from measuring responses to assessing coherence and factuality, rendering earlier benchmarks obsolete [27]. Evaluation must remain adaptive and grounded in current operational contexts.

Table 1. A summary of the Generative AI case studies presented in this paper. Case studies marked with a * are in operation.

| Case Study | Domain | License | Architecture | Organisation |
|---|---|---|---|---|
| Natural Language User Manual Interface* | Manufacturing | Commercial | RAG | Multinational |
| Semantic Document Matching* | Education | Commercial | Document Pipeline | Startup |
| Test Data Generation | Research | Opensource | Agentic | University |

## 4 Case Studies

To illustrate how the principles of the Physics of AI manifest in practice, three separate case studies were selected based on the diversity of sectors and size of the projects. Each of the case studies was conducted under the supervision of the first author. Case Study 1 and Case Study 2 were contract research projects for industry partners and Case Study 3 was a research project. See Table 1 for a summary of the case studies.

### 4.1 Case Study 1: Natural Language User Manual Interface

**Description of use case:** The first case study comes from a project with a multinational motorbike manufacturer. Due to commercial sensitive work the partner prefers not to be named. This project involved collaborating across timezones and involved helping the organisation explore applications of AI in their context. For a preliminary investigation into AI, the partner wanted to support service staff in finding relevant information from service manuals.

**Application of AI:** The architecture chosen for this project was Retrieval Augmented Generation (RAG). As a result, AI was used in two separate places, for embedding content into a database for semantic retrieval and for synthesising an answer to a query. The key to this project was being able to customise the information retrieval pipeline to a acceptable level of performance.

### 4.2 Case Study 2: Semantic Document Matching

**Description of use case:** The second case study involved a project for a small company, Red Velvet AI[1]. This company is looking to build on top of their previous success in the education sector and expand into the higher education sector with an AI powered solution. The core problem for the project was to be able to semantically compare documents. That is, to compare documents based on what is covered in the material independent of the format or structure. For example, given a students transcripts and a collection of units for a course, what prior credit is the student ineligible/eligible for.

**Application of AI:** AI was used to extract content from the documents using OCR technology, to process and understand what the structure of the document is, and to perform the semantic matching. Two areas were particularly critical. First, since the OCR system relies on an external API, the OCR engine itself must perform exceptionally well as there were few viable alternatives if its accuracy was insufficient (building and maintaining a custom model would be prohibitively expensive). Second, it was essential that the document-matching process captures all key attributes to ensure reliable results.

[1]https://www.theredvelvet.ai/

### 4.3 Case Study 3: Test Data Generation

**Description of use case:** This project was a research project used to create the RAGProbe tool. RAGProbe is a tool to be used by developers to generate test data for a RAG pipeline and involves generating question and answer pairs from a corpus of documents. The key approach was to provide test data based on the domain of the problem rather than benchmark dataset or generic question and answers.

**Application of AI:** As the core task was to expose limitations in RAG pipelines, RAGProbe had to generate high quality test data. Large Language Models (LLMs) were used with extensive prompt engineering to generate quality question and answers and to validate the output. The goal of the tool was to create only a few examples that could uncover system flaws, rather than to produce large amounts of data. As such, there were no restrictions on the frequency of LLM calls to create or validate outputs (i.e. generous agentic loops were permitted).

## 5 Case Findings

For each case study, we tracked experiment logs and recorded structured field notes [23]. The collected data is available online[2]. Due to space, in this paper we focus on presenting the key findings from the case studies and interpret them through the lens of the Physics of AI. We reveal how the Physics of AI provides a unifying vocabulary for diagnosing problems, explaining trade-offs, and guiding future design actions.

***CF1. Data validation is essential regardless of the source.*** Across all case studies we have found that labelled data is always approximate no matter the source from: a) subject matter experts, b) generated or, c) application users. For example in Case Study 2, we found that some expert labelled sections appeared in multiple manuals, causing conflicts during an optimisation experiment.

*Interpretation*: The principle of **error is inherent** implies that labelled data from experts, users, or generative sources are only approximations. Their meanings shift interact with software and social processes (**entanglement**). Conflicts, such as duplicate manual sections, are therefore not bugs but emergent properties of socio-technical data.

***CF2. Parameters of AI APIs hinders comparability.*** In Case Study 1, we found that AI APIs have different pricing schemes based on dedicated resources to be purchased by the hour or limited by inputs/outputs. For example, on Microsoft Azure[3] the AI model Phi 4 requires dedicated resources to operate whereas GPT 4o-mini is rate limited based on tokens. We also found not all models or API features are available in every geography due to phased rollouts or regulatory reasons[4].

*Interpretation*: The principles of **resource boundaries** and **training ≠ deployment** highlight that each experiment operates under different constraints. Rate limits, geographic rollout, and pricing tiers mean that two nominally identical experiments can run under divergent conditions.

***CF3. Stakeholder requirements do not fully align with evaluation metrics.*** Automated metrics (e.g. similarity scores, precision/recall, BLEU, etc.) are useful proxies but do not fully capture if the requirement have been achieved. A high embedding similarity score between two items doesn't guarantee they are actually equivalent in meaning or usefulness. It only indicates the model thinks they're close.

*Interpretation*: From the perspective of **partial alignment** and **benchmark expiry**, metrics such as BLEU or cosine similarity serve only as temporary proxies for intent. Their validity diminishes



[2] https://doi.org/10.26187/deakin.30434194
[3] https://azure.microsoft.com/en-au/
[4] https://learn.microsoft.com/en-us/azure/ai-foundry/openai/quotas-limits

as data distributions and user needs evolve. **Error is inherent** reminds us that quantitative performance cannot substitute for human value judgments.

***CF4. Vendor alignment outweighs cost considerations.*** The rapid pace of advancement and investment in AI companies has led to AI APIs becoming both cheaper and more capable over time. However, despite falling prices for users, cost is not the sole factor in enterprise adoption. A 2024 industry survey found that only 1% of enterprise AI leaders cited cost as a primary selection criterion, while ecosystem alignment and vendor trust dominated decision making [25]. Consistent with this, both industry case studies showed that our partners preferred APIs provided by their existing cloud vendors, even when cheaper alternatives were available.

*Interpretation*: The **resource boundaries** and **training ≠ deployment** principles illustrate that practical feasibility often outweighs price. Cloud vendor ecosystems constrain what models can be used due to compliance, data residency, and security considerations. Cheaper options may violate these boundaries and thus be non-deployable.

***CF5. Performance improves by using larger or newer models.*** Case Study 1 and Case Study 3 both used AI models to semantically describe content and create a numerical representation (i.e. an embedding). In our context of embedding-based retrieval, using the largest and latest embedding model yielded superior results compared to smaller models with added complexity (like a re-ranker). Recent research confirms that a strong dense retriever (A large pretrained embedding model) can outperform more complex re-rankers on retrieval tasks [9].

*Interpretation*: The principles of **benchmark expiry** and **training ≠ deployment** show that performance improvements from newer models may invalidate earlier baselines. Upgrading models shifts the efficiency frontier but alters resource profiles and costs. Time-boxed comparative evaluations and re-baselining after major model releases prevent outdated conclusions.

***CF6. Offline evaluations create slow feedback loops.*** When evaluation of an experiment depends on an offline process (such as an expert manually checking results), the feedback to the engineering team is delayed. An engineer might only learn that a experiment failed days or weeks later, once an assessor has reviewed the outputs. This slower cycle contrasts with the fast feedback in traditional software engineering (where tests either pass or fail immediately).

*Interpretation*: According to **human feedback is substrate** expert review remains indispensable but time-consuming. By the time results are returned, underlying data or APIs may have changed, limiting feedback value.

***CF7. Positive practices during experimentation are necessary.*** Traditional software engineering provides frequent rewards (feature shipped to production, code merged, positive code reviews, etc.) that reinforce developer motivation. In contrast, AI experimentation can involve many false starts or "failed" experiments that never see the light of day, which can be demoralising if not managed well.

*Interpretation*: **Interpretability is partial**, **entanglement**, and **human feedback is substrate** clarify that many experiments fail silently. Without structured reflection, such failures erode learning and morale.

## 6 Implications

The case findings demonstrate that the Physics of AI impose unavoidable limits on all AI systems. What remains uncertain in practice is how to contend with these constraints. The following implications (partially) answer this question.

Table 2. Implications for engineering teams based on findings from the three case studies with links to the Physics of AI.

| Implications for Engineering Teams | Findings | Constraints from Physics of AI |
|---|---|---|
| **Invest in Experiment Tracking Infrastructure** | CF1, CF2 | Robustness fragility; Entanglement; Resource boundaries; Training ≠ Deployment |
| **Logging & Observability by Default** | CF5, CF6 | Error is inherent; Partial alignment; Benchmark expiry; Human feedback is substrate |
| **Foster a Learning-Centred Experimentation Culture** | CF3, CF7 | Human feedback is substrate; Entanglement; Interpretability is partial |
| **Define a Business Metric for the Problem** | CF3, CF4 | Partial alignment; Training ≠ Deployment; Error is inherent; Benchmark expiry; Resource boundaries; Human feedback is substrate; Interpretability is partial |
| **Know the Bias of your Evaluators** | CF1, CF2 | Error is inherent; Entanglement; Benchmark expiry; Training ≠ Deployment; Partial alignment |

**Invest in Experiment Tracking Infrastructure.** Experimentation is key to discover if AI provides value in a given context. Experimentation infrastructure solves the challenges with reproducitibilty and enables to teams to discover how the Adaptation constraints (Section 3.1) impact the system. This practice mitigates the comparability gaps highlighted in **CF2**, and helps with continuous data validation (**CF1**).

**Logging & Observability by Default.** Experimentation alone is insufficient. Teams also need observable systems with appropriate logging in both development and production systems. Logging helps address all three of the constraint types (Adaptation, Regulation and Resilience) through insight into what is happening at the system level. Observable systems shorten the slow review loops (**CF6**) and make it easier to distinguish which models perform best (**CF5**).

**Foster a Learning-centred Experimentation Culture.** Discovering the workflow and technology pipeline that generates value requires experimentation, not just engineering. This involves people being willing to experiment and change the way development tasks are approached. By focusing on the team, awareness of all constraints is formed and provides a foundation for new application specific approaches to address the limitations. A learning centred culture helps with recognition (**CF7**), and aligning with stakeholder goals (**CF3**).

**Define Business Metrics for the Problem.** A business metric is a metric that the business wants to see improve from the adoption of AI. Defining a business metric ensures that success is grounded in user- or assessor-aligned objectives, not abstract numerical improvements. Numerical metrics (i.e. accuracy, f1-score) are proxies for a business metric. This practice focuses on helping with the Regulation (Section 3.2) and Resilience constraints (Section 3.3). This addresses an over-reliance on quantitative proxies (**CF3**) and focusing on cost alone (**CF4**).

**Know the Bias of your Evaluators.** All evaluators, both human annotators and other AI models (i.e. LLM-as-a-judge), are biased. Makes sure effort is invested in knowing the bias of each evaluator and have multiple evaluators that reduce the bias overall. Strategies include repeated sampling, bias correction, and hybrid human–AI calibration approaches. This practice: a) simplifies

comparability across models (**CF2**), and b) helps validate data sources (**CF1**).

> **The Silver Lining:** The Physics of AI impacts the engineering methodology and practices, not necessarily the outcome.

## 7 Conclusion

This paper introduced the Physics of AI as a unifying framework that articulates the nine structural limits governing all artificial intelligence systems. Across three generative AI case studies, we demonstrated how these principles manifest in practice. The findings reveal that while these limits cannot be eliminated, they can be managed through disciplined experimentation, observability, and governance. The resulting framework shifts the focus from idealised optimisation to value generation under constraint. Practical actions such as investing in experiment tracking, embedding observability, defining business-grounded metrics, and understanding evaluator bias translate these principles into engineering practice. Future research should investigate how the Physics of AI can inform cross-organisational governance, adaptive benchmarking, and regulation, helping to establish AI engineering as a discipline grounded in its own physical-like laws.



## Acknowledgments

To our industry partners including Red Velvet AI and to our team that made this projects possible: Winnie Tong, Jon Tse, Zac Brannelley, Nhat Duong and Taylan Selvi. Generative AI contributed to the writing process by helping shape iterations and improve structure.

## References

[1] Sherif Akoush et al. 2022. Desiderata for Next Generation of ML Model Serving. *arXiv preprint arXiv:2210.14665* (Nov 2022).

[2] Rob Ashmore et al. 2021. Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. *ACM computing surveys (CSUR)* 54, 5 (May 2021), 1–39.

[3] Sourav Banerjee et al. 2025. LLMs will Always Hallucinate, and We need to Live with This. In *Intelligent Systems Conference*. Springer, Amsterdam, The Netherlands, 624–648.

[4] Charles H Bennett and Rolf Landauer. 1985. The Fundamental Physical Limits of Computation. *Scientific American* 253, 1 (July 1985), 48–57.

[5] Houssem Ben Braiek and Foutse Khomh. 2025. Machine Learning Robustness: A Primer. In *Trustworthy AI in Medical Imaging*. Academic Press, 37–71.

[6] Ben Cottier et al. 2025. The Rising Costs of Training Frontier AI Models. *arXiv preprint arXiv:2405.21015* (Feb 2025).

[7] El-Mahdi El-Mhamdi and Lê-Nguyên Hoang. 2024. On Goodhart's Law, with an Application to Value Alignment. *arXiv preprint arXiv:2410.09638* (Oct 2024).

[8] Mikhail Evtikhiev et al. 2023. Out of the BLEU: How Should We Assess Quality of the Code Generation Models? *Journal of Systems and Software* 203 (Sep 2023), 111741.

[9] Marie Al Ghossein et al. 2024. ICLERB: In-Context Learning Embedding and Reranker Benchmark. *arXiv preprint arXiv:2411.18947* (Nov 2024).

[10] Ian J Goodfellow et al. 2015. Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572* (Mar 2015).

[11] Xingquan Guan and Henry Burton. 2022. Bias-variance Tradeoff in Machine Learning: Theoretical Formulation and Implications to Structural Engineering Applications. *Structures* 46 (Dec 2022), 17–30.

[12] Thilo Hagendorff. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and machines* 30, 1 (Feb 2020), 99–120.

[13] Ronan Hamon et al. 2020. Robustness and Explainability of Artificial Intelligence. *Publications Office of the European Union* 207, 40 (Jan 2020).

[14] Jean-Marie John-Mathews. 2022. Some Critical and Ethical Perspectives on the Empirical Turn of AI Interpretability. *Technological Forecasting and Social Change* 174 (Jan 2022), 121209.

[15] Lisa M Koch et al. 2024. Distribution Shift Detection for the Postmarket Surveillance of Medical AI Algorithms: a Retrospective Simulation Study. *NPJ Digital Medicine* 7, 1 (May 2024), 120.
[16] Pantelis Linardatos et al. 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23, 1 (Dec 2020), 18.
[17] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is both Important and Slippery. *Queue* 16, 3 (Jun 2018), 31–57.
[18] Jie Lu et al. 2018. Learning under Concept Drift: A Review. *IEEE transactions on knowledge and data engineering* 31, 12 (Oct 2018), 2346–2363.
[19] Sean McGregor. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 15458–15463.
[20] Brent Mittelstadt. 2019. Principles Alone Cannot Guarantee Ethical AI. *Nature machine intelligence* 1, 11 (Nov 2019), 501–507.
[21] Anka Reuel et al. 2024. Betterbench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. In *Advances in Neural Information Processing Systems*, Vol. 37. Vancouver, Canada, 21763–21813.
[22] David Sculley et al. 2015. Hidden Technical Debt in Machine Learning Systems. In *Advances in Neural Information Processing Systems*, Vol. 28. Montreal, Canada.
[23] Carolyn B. Seaman. 1999. Qualitative Methods in Empirical Studies of Software Engineering. *IEEE Transactions on Software Engineering* 25, 4 (Aug 1999), 557–572.
[24] Rachel L Thomas and David Uminsky. 2022. Reliance on Metrics is a Fundamental Challenge for AI. *Patterns* 3, 5 (May 2022).
[25] Tim Tully et al. 2024. The State of Generative AI in the Enterprise. https://menlovc.com/2024-the-state-of-generative-ai-in-the-enterprise/. [Accessed 28-09-2025].
[26] Yixin Wang and Michael I Jordan. 2024. Desiderata for Representation Learning: A Causal Perspective. *Journal of Machine Learning Research* 25, 275 (Aug 2024), 1–65.
[27] Yi-Ting Yeh et al. 2021. A Comprehensive Assessment of Dialog Evaluation Metrics. *arXiv preprint arXiv:2106.03706* (Jul 2021).